GO-ESSP Meeting, RAL, June 2005

The BODC Parameter Markup and Usage Vocabulary Semantic Model

Roy Lowry

British Oceanographic Data Centre

Presentation Outline

- Parameter codes and their metadata load
- EnParDis Project
- BODC PMUV Semantic Model
- Issues: Synonyms and Tooling
- Points to Ponder

What is a Parameter Code?

- According to the original oceanographic data standard (GF3) a parameter code is a key attached to a data value that:
  - Specifies:
    - What was measured
    - How it was measured
    - Actual (not canonical) units of measurement
  - Is defined through the attributes of a parameter dictionary
  - Includes semantics (e.g. TEMP7RTD, TEMP7STD)
- The data models of IODE data centres were strongly influenced by GF3 concepts (even if the format was little used)
- Parameter codes are therefore endemic in legacy oceanographic data

What is a Parameter Code?

- Parameter codes have been grossly abused by the oceanographic data management community
- Codes mapped to free text fields causing
  - Compromised semantic purity (e.g. vague spatio-temporal co-ordinates like ‘sea-surface’ introduced)
  - Incomplete or ambiguous specifications. Sample GF3 parameters:
    - Sea temperature (estuaries?)
    - Potential temperature (of what?)
    - Potential air temperature (OK!)
    - Wet bulb temperature (could be in water!)
  - Metadata overload (e.g. taxon names included in parameter – what was measured - descriptions)
  - Random scatterings of synonyms
  - Parameter semantics in unit definitions (e.g. per gram dry weight)

What is a Parameter Code?

- BODC joined in this abuse through a Parameter Dictionary following the GF3 model
- BODC Parameter Dictionary originally mapped code to:
  - Two plain-text fields of what measured (parameter) and how (parameter subgroup)
  - Units specification
  - Valid data range
  - Formatting information
  - Abbreviated description label
- As chemistry and biology were added the plain-text fields evolved into a total mess

Enabling Parameter Discovery (EnParDis)

- EnParDis was a one-off injection of NERC funding aiming to:
  - Integrate taxonomic knowledge into the BODC Parameter Dictionary
    - Include data on penguins in a query for ‘birds’
    - Based on ITIS and works providing taxa are in ITIS
  - Totally overhaul the parameter plaintext descriptions
    - Standardise terms and syntax
    - Eliminate implied semantics
- Plaintext field overhaul achieved by using structured text (concatenated elements from a semantic model)

Semantic Model

- The Semantic Model maps each parameter code to a set of atomic metadata elements populated by entries from controlled vocabularies
- The parameter description is built by structured concatenation of the Semantic Model elements
- The Semantic Model forms a flexible interface between legacy systems based on parameter codes and/or semantically poor text descriptions and modern (meta)data content models
- The Parameter Dictionary becomes a registry of valid Semantic Model element combinations (insurance against the introduction of the green dog)

Semantic Model

- The parameter description is built up as three themes:
  - What theme – what was measured
  - Where theme – where it was measured (sphere NOT spatio-temporal co-ordinates or their textual representation like sea surface)
  - How theme – how it was measured

- Example
  - Temperature (What)
  - of the water column (Where)
  - by CTD (How)

What Theme

- Parameter Entity of Measurand Entity (chemical, physical or biological) by Measurand Entity
- Examples
  - Clearance rate of Dinophycae by Acartia
  - Concentration of nitrate+nitrite
  - Temperature

Parameter Entity

- Three Semantic Elements
  - Parameter Name
    - Example: concentration
  - Parameter Statistic
    - Example: standard deviation
  - Parameter Subgroup
    - Example: v/v

Biological Entity

- **Nine Semantic Elements**
  - Taxon Name
  - ITIS code for taxon
  - Taxon size
  - Taxon gender
  - Taxon development stage
  - Taxon morphology (shape terms)
  - Taxon subcomponent (body parts)
  - Taxon colour
  - Taxon subgroup (subdivision ‘bucket’)

Chemical Entity

- Two Semantic Elements
  - Chemical name
    - Example: carbon
  - Chemical subgroup
    - Example: organic

Physical Entity

- **Three Semantic Elements**
  - **Physical Name**
    - **Examples: sea surface elevation, temperature**
  - **Physical subgroup**
    - **Example: IPTS-68**
  - **Datum**
    - **Example: Ordnance Datum Newlyn**

Where Theme

- Relationship and a Sphere Entity or Biological Entity
- Examples
  - Per unit volume of the water column
  - Of the atmosphere
  - Per unit wet weight of Mytilus edulis flesh
  - Per unit dry weight of sediment
  - Per unit area of the water column

Sphere Entity

- **Four Semantic Elements**
  - **Sphere Name**
    - **Example: sediment**
  - **Sphere subgroup**
    - **Example:<63um**
  - **Sphere phase**
    - **Examples: particulate, aerosol, gaseous, dissolved plus reactive particulate**
  - **Sphere phase subgroup**
    - **Examples: >GF/F, 2-20um**

How Theme

- **Sample processing entity**
  - **Example: radiotracer inoculation and incubation in natural sunlight**
- **Analysis entity**
  - **Example: proportional counting**
- **Data processing entity**
  - **Example: conversion to carbon using unspecified algorithm**

How Theme

- Sampling and Data Processing Entities are single entities
- Analysis Entity has two Semantic Elements
  - Analysis Description
  - Analysis Instance Discriminator (multiple sensors)

Issues

- Synonyms
  - What synonyms do we need?
  - How do we store them?
  - How do we utilise them?
- Tooling
  - Tools for code assignment
  - Tools for dictionary expansion

Synonyms

- **Synonyms required for:**
  - What Theme
  - Parameter Name
  - Parameter Entity
  - Chemical Name
  - Chemical Entity
  - Physical Name
  - Physical Entity
  - Biological Entity
  - Taxon name
  - Parameter Description
  - Probably more as well

Synonyms

- The following information needs to be known for each synonym
  - The Semantic Model entity type
  - The primary term
  - The secondary term
- Could be managed through a conventional relational schema, but RDF seems more attractive as there is more to relationships than ‘synonymous’
- Current thinking on synonym exposure is to produce multiple parameter descriptions for a single parameter code incorporating all synonym combinations

Tooling

- Web services to give access to code definitions, model elements, controlled vocabularies, mappings and synonyms
- Automated data markup tooling based on model semantic element specification
- Automated request mechanism for dictionary population extension with efficient moderation mechanism
- Could be configured as a single tool sitting on a common set of services

Points to Ponder

- What is the mapping between components of the Semantic Model and a CF Standard Name?
- How can the CF Standard Name list and the BODC Data Markup vocabulary be integrated into a unified resource covering both the oceanographic and atmospheric domains?